Comparative analysis of some search engines

Authors:
Joseph Edosomwan[1]
Taiwo O. Edosomwan[1]

Affiliation:
[1]College of Education, Ekiadolor-Benin, Edo State, Nigeria

Correspondence to:
Joseph Edosomwan

email:
maryjoe872002@yahoo.com

Postal address:
College of Education, Ekiadolor-Benin, PMB 1144, Benin City, Edo State, Nigeria

Keywords:
catalogue; crawler; index; precision; response time; search engine; spider

Dates:
Received: 14 May 2009
Accepted: 26 May 2010
Published: 29 Oct. 2010

How to cite this article:
Edosomwan J, Edosomwan TO. Comparative analysis of some search engines. S Afr J Sci. 2010;106(11/12), Art. #169, 4 pages. DOI: 10.4102/sajs.v106i11/12.169

This article is available at:
http://www.sajs.co.za

ABSTRACT

We compared the information retrieval performances of some popular search engines (namely, Google, Yahoo, AlltheWeb, Gigablast, Zworks and AltaVista and Bing/MSN) in response to a list of ten queries, varying in complexity. These queries were run on each search engine and the precision and response time of the retrieved results were recorded. The first ten documents on each retrieval output were evaluated as being 'relevant' or 'non-relevant' for evaluation of the search engine's precision. To evaluate response time, normalised recall ratios were calculated at various cut-off points for each query and search engine. This study shows that Google appears to be the best search engine in terms of both average precision (70%) and average response time (2 s). Gigablast and AlltheWeb performed the worst overall in this study.

INTRODUCTION

Searching on the World Wide Web has become a part of our daily life as the Web is now a necessary tool for collecting information and, undoubtedly, it provides convenience in information retrieval because it can combine information from many different websites.[1] Akeredolu[2] officially described the Web as 'a wide-area hypermedia information retrieval initiative aimi ng to give universal access to a large universe of documents'. In simpler terms, the Web is an Internet-based computer network that allows users on one computer to access information stored on another through the world-wide network.[3]

The ultimate goal in designing and publishing a webpage is to share information. However, the high number of webpages added to the Web daily has made the Web a sea of all kinds of data and information, which provides a challenge for information retrieval. The amount of information on the Web, as well as the number of hosts and domain names registered worldwide, are growing rapidly.[4] There are currently more than 1 trillion webpages and it is estimated that the number of webpages will continue to grow to infinity. Several billion webpages are added to the Web daily.[4] This new information must be made accessible to everybody for a webpage to achieve its intended goal. To overcome these retrieval problems, more than 20 companies and institutions have developed search tools, such as Yahoo, AltaVista, Google and Lycos.

Search tools can be classified into two basic types: directories and search engines.[5] The main difference between directories and search engines is that a directory is built by people, whereas the search engines database is created by software known as spiders or robots. Searching, instead of browsing, is the main feature of search engines. The advantage of search engines over directories is that they are very comprehensive, often including thousands of sites in the results listed. The disadvantage therefore is having to weed through thousands of irrelevant sites to find what you are looking for, because, although search engines attempt to list sites in order of relevance, this relevance is determined by a mathematical formula that is far from perfect. Search engines are particularly useful when searching for a specific topic that may not be found in a directory.[6] Search engines are tools for searching for information and directories are collections of human-reviewed Web sites that have been arranged into topical categories. Therefore directories impact search engines, hence the interchangeable use. Although they have different search strategies, both search engines and directories have similar interfaces and are commonly known as search engines; we therefore refer to both types as search engines in this paper.

Search engines are listed among the most accessed sites. Search engines create and maintain an index of words within documents on the Web. They return to a user a ranked list of relevant documents as search results. A few of the results may be valuable to a user whilst the majority usually is irrelevant.[7]

As the Web continues to grow, most search engines are faced with serious challenges; the Web is growing much faster than any present technology can possibly index and so the results may become out-of-date. Many webpages are updated frequently, which forces the search engines to visit them periodically. Many dynamically generated sites are also not indexable by search engines. The largest search engines (i.e. those with the largest indexes) have done an impressive job in extending their reach, but the technology has had to scale dramatically to keep up with the growth of the Web. In 1994, the World Wide Web Worm, one of the first Web search engines, had an index of 110 000 webpages and Web accessible documents.[2] As at 2006, Google had indexed 25 billion webpages; presently, in 2010, Google indexes 19.2 billion webpages each day.[8]

At the same time as the number of webpages is increasing, the number of queries that search engines are required to handle has also grown incredibly. Currently, Google receives an average of 400 million queries per day and as of July 2010,[9] AltaVista claimed it handled about 13 million queries per day.[10]

Considering the challenges faced by most search engines, the need for better search engines to more easily and quickly locate relevant information to meet the various needs of Web users has become increasingly important. The right choice of search engine helps to reduce the difficulties encountered in the retrieval of information from the Web. Faced with the option of so many search engines, users can be easily confused.

Users tend to always return to one or two search engines with which they are comfortable. However, which search engine actually satisfies a user's need and which is the best? To answer these questions, a user must clarify their needs and which features they prefer, for example the amount of information retrieved, the speed at which search results are retrieved or the relevance of the search results.

This study employed an empirical approach to evaluate the precision and speed of information retrieval of some selected search engines. Our results will allow users to have a better understanding of a search engine's capabilities, make inferences about different search engines and discover avenues for further research.[11]

How do search engines work?

There are different ways to organise Web content but every search engine has the same basic parts which include a crawler or spider, an index or catalogue, and an interface or query module (Box 1). Users enter a search term through a predefined query module, specific to each search engine. Typically, the search engine works by sending out a spider to fetch as many documents as possible. Then another program called an indexer reads these documents and creates an index based on the words contained in each document. Each search engine also uses a proprietary algorithm to create indices which ideally enable only meaningful results to be returned for each query. In general, a search engine starts with a set of predefined Web addresses and downloads them. For each page, it extracts the uniform resource locator or URL in order to follow them later in a specified manner. It then indexes all of the words and phrases as well as the relative position of the words to each other. The user can then search this index through the results retrieved for the presence of a particular word, phrase or combination of words in a Web document.

BOX 1
Definition of terms

Search engine: A program that searches documents for specified keywords and returns a list of the documents where the keywords were found, ranked in order of relevance. It allows one to ask for content meeting specific criteria (typically those containing a given word or phrase) and retrieves a list of references that match those criteria.[12]

Directory: A manual catalogue of sites on the Internet. People create categories and assign sites to a place within a structured index. An example of a typical directory is Yahoo, which screens all relevant information and assigns this information to an address. Yahoo also orders sites so that the most relevant or comprehensive in each category appears first on the list. This search feature can help people quickly find targeted information on more general topics.[13]

Crawler/Spider: Visits webpages following links, updating pages and adding new pages when it comes across them.[14]

Index/Catalogue: Where a spider's collected data is stored i.e. it contains a copy of every webpage that the spider finds.

Query: The keyword or question entered by the user requesting the search engine to search for.

Response time: The period between issuing a search query and the display of the first search results.[15]

Precision: The relevance of a search result to a search query.

METHODS

Ten search queries were used to test seven different search engines; both the precision and response time of the search results retrieved were then compared amongst the search engines.

Selection of search engines

The search engines selected for comparison in this study were Yahoo, Google, Gigablast, AlltheWeb, Zworks, AltaVista and Bing/MSN.

During the process of selecting Web search engines to be evaluated, attention was paid to including a diverse range of search engines so that the results obtained could serve as a basis for evaluating the search algorithm used by the various search engines. Some of the selected search engines are not the most popular or most familiar. The results of the study will therefore enlighten users about their different capabilities and thereby potentially increase the usage of the better performing search engines.

Many search engines also index resources stored on other Internet applications, such as discussion groups and Gopher (a network that directs users to companies providing certain products and/or services), in addition to Web information; this study however only considered Web databases. Unified Web search engines such as CUSI (Configurable Unified Search Index) also were not considered because they only compile existing Web information and do not provide anything new.[16]

Test queries

Ten search queries were designed for use on all of the search engines. These queries were designed to test various features that each search engine claims to have, as well as to represent different levels of searching complexity. The queries also were designed to fall within the domain of Information Technology for the purpose of familiarity, such that the investigators could judge the search results for relevance. The ten queries were classified into four groups as follows:

TABLE 1A
Response time, measured in seconds, of the search engines to four selected queries, varying in complexity, during off-peak hours

Query	Yahoo	Google	Gigablast	AlltheWeb	Zworks	AltaVista	Bing/MSN	Mean	s.d.
1	6	2	3	9	5	6	2	5	3
6	10	2	2	10	5	5	4	5	3
8	8	2	7	6	5	10	2	6	3
10	7	3	8	9	4	6	3	6	2
Mean	8	2	5	9	5	7	3	-	-
s.d.	2	1	3	2	1	2	1	-	-

s.d., standard deviation

TABLE 1B
Response time, measured in seconds, of the search engines to four selected queries, varying in complexity, during peak hours

Query	Yahoo	Google	Gigablast	AlltheWeb	Zworks	AltaVista	Bing/MSN	Mean	s.d.
1	38	12	18	25	9	15	17	19	8
6	12	9	25	38	17	27	15	21	8
8	25	23	18	24	13	21	15	19	4
10	23	18	17	32	24	25	14	21	4
Mean	25	16	20	30	16	22	15	-	-
s.d.	9	5	3	6	6	5	1	-	-

s.d., standard deviation

Article #169
South African Journal of Science

TABLE 2
Precision scores* for each query performed on each search engine

Query	Yahoo	Google	Gigablast	AlltheWeb	Zworks	AltaVista	Bing/MSN	Mean
1	0.8	0.7	0.7	0.6	0.6	0.7	0.5	0.7
2	0.9	0.6	0.6	0.6	0.3	0.6	0.7	0.6
3	0.7	0.8	0.3	0.8	0.7	0.6	0.8	0.7
4	0.3	0.8	0.5	0.5	0.6	0.5	0.5	0.5
5	0.6	0.8	0.6	0.7	0.5	0.5	0.6	0.6
6	0.7	0.8	0.4	0.5	0.5	0.7	0.6	0.6
7	0.7	0.9	0.3	0.6	0.5	0.6	0.5	0.6
8	0.5	0.6	0.5	0.5	0.3	0.4	0.6	0.5
9	0.5	0.5	0.2	0.3	0.4	0.3	0.3	0.4
10	0.4	0.4	0.2	0.3	0.2	0.1	0.3	0.3
Mean	0.6	0.7	0.4	0.5	0.5	0.5	0.5	-
s.d.	0.6	0.5	0.6	0.6	0.5	0.5	0.5	-

*Precision was calculated as a value between 0 and 1, with 1 representing ten out of ten search results being relevant
s.d., standard deviation

TABLE 3
Ranking of the search engines according to their response times, precision scores and overall performances

Criteria	Search engine						
	Yahoo	Google	Gigablast	AlltheWeb	Zworks	AltaVista	Bing/MSN
Response time	6	1	3	7	3	5	2
Precision	2	1	7	3	3	3	3
Mean rank	4	1	5	5	3	4	2.5
Ranking*	4th	1st	7th	7th	3rd	4th	2nd

*Overall performance ranking based on mean ranking

A. Short queries:
- What is data mining? (Query 1)
- Web browsers (Query 2)
- Neural network (Query 3)
- Evolution of microprocessor (Query 4)
- Keyword surfing (Query 5)

B. Boolean logic (AND/OR) queries:
- Searching AND sorting (Query 6)
- Clustering OR clustering algorithm (Query 7)

C. Natural language queries:
- Search the Internet using natural language (Query 8)
- How do I get the best search result on the Web? (Query 9)

D. Long query:
- I found a cool webpage but I lost it. How do I get it back? (Query 10)

For each query, only the first ten search results were evaluated. For most users, the first ten retrieved results are the most important, i.e. almost all users hope that the first ten search results will provide what they are looking for and if this is not the case, they become frustrated and usually try another search engine.[17] Considering that all selected search engines display results in descending order of relevance, it is believed that this methodology did not critically affect the validity of the results.

Test environment

Microsoft Internet Explorer was chosen as the Web browser for the study because it is compatible with all the search engines selected and is the most widely used browser locally. Two computers with different configurations but with the same parameters were used: an Acer computer with an Intel Celeron M Processor 440, 80 GB hard disk (1.86 GHz speed) and 52 MB DDR2 memory and a Hewlett Packard computer (2.10 MHz speed) with an AMD Semipro SI-42 processor, 140 GB hard disk and 1 GB RAM. One computer was used for the entire experiment, which was repeated for validity on the second computer, i.e. each query was run twice. The results shown are those obtained from the Hewlett Packard computer. Results from the repeated exercise are not presented because they were comparable and do not alter the outcomes of the study.

Ideally, each query should be executed on all search engines at the same time, so that if a relevant page is added, none should have an advantage of being able to index the new page over the other. For this study, that was not practically possible and so each query was searched on all the search engines within thirty minutes of each other on the same day. Those search engines returning an error of '404' (i.e. path not found) or '603' (i.e. server not responding) were noted in order to be returned to. Return visits were made at different times of the day to allow for the possibility that the site might have a regular down time for maintenance.

Response time

Response time was calculated as the period between entering a search query and retrieval of the first search results and was measured by a stopwatch. We selected one query from each group to assess response time. The queries selected were: Query 1 (Group A), Query 6 (Group B), Query 8 (Group C) and Query 10 (Group D). The average response times for each search engine and for each selected query were then calculated.

Precision

For this study, precision was defined as the relevance of a search result to a search query and was determined separately by both investigators for the first ten search results. We checked the content of each retrieved result to determine whether it satisfied the expected result, but did not attempt to read the full-text Web document by following the links provided because of time considerations and variable reliability of the links. A precision score was calculated based on the number of results within the first ten retrieved deemed to be relevant (i.e. a score of 1 indicates that all ten search results were relevant and a score of 0.5 indicates that only five of the first ten results were relevant). In order to assess the overall performance of each search engine we evaluated, we not only computed the average precision score for each query, but also calculated the average precision score, based on all ten queries, for each search engine.

RESULTS AND DISCUSSION

Response time

The mean response times for all the search engines were within the range of 2 s – 9 s during off-peak hours. During peak hours,

mean response time increased to 15 s and went as high as 30 s. The individual and mean response times for each search engine and for each query during off-peak and peak hours are shown in Tables 1A and 1B, respectively.

Precision

The precision score for each query on each search engine is tabulated in Table 2. The mean precision scores for each search engine ranged from 0.4 to 0.7. Although the ranking of the precision scores varied amongst the search engines depending on the query, Google obtained the highest mean precision score of 0.7 while Yahoo obtained the second highest precision score of 0.6. Gigablast obtained the lowest precision score of 0.4.

The highest precision score for query 10 (i.e. a long query) was 0.4 (Google and Yahoo), which indicates that the search engines had more difficulty in processing long queries compared to the shorter queries. This result implies that users wanting the most relevant search results should be as precise as possible in their search queries, supplying only the most important terms.

Overall performance

Table 3 shows the seven search engines ranked in terms of their response times (from shortest to longest) and precision scores (from highest to lowest), with a rank of 1 denoting the best performer. The average of both rankings gives an indication of the overall performance of each search engine.

CONCLUSION

For both response time and precision, Google proved to be the best performer of all the search engines evaluated. Hence it is the search engine we recommend. MSN/Bing, the second best performer, is also recommended. Gigablast and AlltheWeb were the worst overall performers in this study.

REFERENCES

1. Waheed IB, Coop L, Kogan M. Integrated pest management (IPM) and Internet-based information delivery systems. Neotrop Entomol. 2003;32(3):373–383.
2. Akeredolu GF. Internet search tools: A comparative study and analysis. MSc thesis, Ibadan, University of Ibadan, 2005.
3. Lancaster FW, Fayen EG. Information retrieval on-line. Los Angeles: Melville Publishing Co; 1973.27. Leighton HV. Performance of four World Wide Web (WWW) index services: InfoSeek, Lycos, WebCrawler, and WWW Worm [homepage on the Internet]. c1995 [cited 2006 Jan 7]. Available from: http://www.curtin.edu.au/curtin/library
4. Alpert J, Hajaj N. We know web was big. Web search infrastructure team. [homepage on the Internet]. c2008 [cited 2008 July 25]. Available from: http://googleblog.blogspot.com/2008/07/we-know-web-was-big.html
5. Liu H, Weber RR. Web crawler [homepage on the Internet]. c2010 [cited 2010 June 7]. Available from: http://en.wikipedia.org/wiki/Web_crawler
6. Silverstein C, Pederson JO. Almost-constant time clustering of arbitrary corpus subsets. Paper presented at: SIGIR '97. Proceedings of the 20th International ACM/SIGIR Conference on Research and Development in Information Retrieval; 1997 July 27–31; Philadelphia, PA. New York: ACM Press; 1997. p. 60–66.
7. Tillman HN. Evaluating quality on the Net [homepage on the Internet]. c2003 [cited 2003 March 28]. Available from: http://www.hopetillman.com/findqual.html
8. Smith N. Google images. Product manager Google images [homepage on the Internet]. c2010 [cited 2010 July 25]. Available from: http://googleblog.blogspot.com/search/label/search
9. Google search [homepage on the Internet]. c2010 [2010 July 20]. Available from: http://en.wikipedia.org/wiki/google_search
10. AltaVista [homepage on the Internet]. c2010 [2010 July 20]. Available from: http://en.wikipedia.org/wiki/Altavista
11. Westera G. Comparison of search engine user interface capabilities [homepage on the Internet]. c2002 [cited 2006 Jan 7]. Available from: http//www.curtin.edu.au/curtin/library/staffpage/gwpersonal
12. Leonard AJ. Where to find anything on the Net [homepage on the Internet]. c1996. [2006 Jan 7]. Available from: http://www.emeraldinsight.com/journals.htm?articlerd
13. Liu K, Yu C, Meng W. Discovering the representative of a search engine. Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM '02); 2002 Nov 4–9; Mclean, Virginia. New York: ACM Press; 2002. p. 558–565.
14. Singh R. Performance of World Wide Web search engines: A comparative study. Vision 2. Knol. [homepage on the Internet]. c2008 [cited 2008 July 27]. Available from: http://knol.google.com/k/performance-of-world-wide-web-search-engines-a-comparative-study#
15. Harman D. Overview of The Second Text Retrieval Conference (TREC-2). Inf Process Manag. 1995;31(3):271–289.
16. Heting C. Information representation and retrieval in the digital age (ASIST Monograph Series). New Jersey: Information Today; 2003.
17. Courtois M, Berry MN. Results ranking in Web search engines. New York: Lukas; 1999.

Article #169 South African Journal of Science